

The University of Chicago

MEASURING COMPETENCE IN
HIGH-SCHOOL ENGLISH

A PART OF A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE
SOCIAL SCIENCES IN CANDIDACY FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

JUNE, 1945

By
VERNA WHITE

Reprinted from

THE SCHOOL REVIEW

Vol. LV, No. 6, June, 1947; Vol. LV, No. 8, October, 1947

The University of Chicago

MEASURING COMPETENCE IN
HIGH-SCHOOL ENGLISH

A PART OF A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE
SOCIAL SCIENCES IN CANDIDACY FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATION

JUNE, 1945

By
VERNA WHITE

Reprinted from
THE SCHOOL REVIEW
Vol. LV, No. 6, June, 1947; Vol. LV, No. 8, October, 1947



MEASURING COMPETENCE IN ENGLISH OF ARMED SERVICES PERSONNEL

VERNA WHITE
Syracuse University

★

THE program of the United States Armed Forces Institute¹ made necessary the construction of an examination to measure the competence in English at the high-school level of men and women in the armed forces. The achievement on this test was to be used as one of the bases for promotion within the various branches of the service and for placement at the proper level in educational institutions upon release from the service. Means were made available for constructing the most valid test possible to distinguish among tenth-, eleventh-twelfth-, and thirteenth-grade students, and a standardization program was planned to give a great variety of data to indicate the validity and reliability of that test. This article describes the process used in constructing the test and presents data on the validity and reliability derived from the standardization program. An article to follow will discuss the extension of the validity study by means of the case-study method applied to one institution.

PRELIMINARY PROCEDURES

The first task in the construction of a test for nation-wide use was to determine what should be tested. When a test is to be used in a single school, the curriculum of that school indicates and delimits the field to be tested and the educational philosophy giving direction to the test. There was no such defined area for an evaluative instrument for administration to men and women from all the schools of the country. Authorities disagreed widely on what constituted a course in high-school English. Therefore the examiners called together a group of consultants² to help define the areas to be tested, to give their pooled judgments of the appropriate emphases, and to suggest certain specifications for the test. They decided that, of the four skills in English (reading, writing, speaking, and listening), only reading and writing could be measured by objective tests that could be scored by

² The consultants were Harold A. Anderson, University of Chicago; John J. De Boer, Chicago Teachers College; Isabel Kincheloe, Curriculum Bureau, Chicago public schools; Henry C. Meckel, University of California; Robert C. Pooley, University of Wisconsin; Rachel Salisbury, Milton College; and Dora V. Smith, University of Minnesota.

¹ Ralph W. Tyler, "Appraisal of Military Training and Experience," *Journal of the American Association of Collegiate Registrars*, XVIII (July, 1943), 345-52.

machines. (The test here discussed measures only the first of these, reading.)

It was concluded at the meeting of the consultants that the type of test which would best account for the great divergences in the English curriculum would be one which measured the wide variety of habits, skills, attitudes, and appreciations which are essential to subsequent literary pursuits. In accordance with this principle, the examiner adopted Gray's definition of reading for general education as the one most appropriate to be used as a basis for specific objectives. Gray states:

Good habits of recognition, comprehension, and speed of reading are not sufficient, however. Of even greater significance today in both school and adult life are the reader's reactions to, and his use of, the ideas apprehended. In this connection he not only recognizes the essential facts or ideas presented but also reflects on their significance, evaluates them critically, discovers relationships between them, and clarifies his understanding of them. According to this conception, reading includes much that psychologists and educators have commonly called thinking. Any concept of reading that fails to include an inquiring attitude, a clear recognition of purpose, orderly reflection, and critical evaluation is inadequate today.³

With the basic concepts established, it was possible to proceed with the

³ William S. Gray, "The Nature of Reading Problems in General Education," *Recent Trends in Reading*, p. 8. Compiled and edited by William S. Gray. Proceedings of the Conference on Reading Held at the University of Chicago, Vol. I. Supplementary Educational Monographs, No. 49. Chicago: University of Chicago Press, 1939.

collection and classification of the behavioral outcomes which are considered desirable for the teaching of reading and interpretation of literature in high school. These objectives were gathered from study of the needs of youth, of the needs of society, and of the consensus of opinion of specialists in the field of English. The objectives were grouped under the following major classifications:

- I. The student applies previously acquired knowledges, skills, and abilities basic to a more complete understanding of his reading.
- II. The student gets the "plain sense" of the material that he reads.
- III. The student reads critically.
- IV. The student goes beyond the written word.
- V. The student reacts to his reading emotionally and imaginatively.
- VI. The student incorporates the manifold results of his reading into his personal development.
- VII. The student becomes acquainted with an increasing number of the books and authors that have made and are making significant contributions to our culture.⁴ [The seventh item was tested in the second part of the test and is not dealt with in this discussion of the construction of the test.]

These objectives were submitted to the consultants for criticism. Their unanimous acceptance of the list, with minor changes, is an indication that valid objectives were chosen.

A survey of the tests in the field of high-school reading and literature and

⁴ For a complete listing of these objectives, see Verna White and J. B. Enochs, "Testing the Reading and Interpretation of Literature," *English Journal*, XXXIII (April, 1944), 172-77.

a study of the criticisms of these tests made evident the lack of an instrument which would fulfil the requirements of the specific situation. Eighteen existing tests examined⁵ failed to meet the criteria set up. Thus, without further analysis, it could be concluded that a new test must be constructed to meet the specifications for this particular testing program.

CONSTRUCTION OF THE TEST

Step 1. Choosing a central theme for each test and selecting passages upon that theme.—Two forms of the test were constructed—one strictly for military use and the other a comparable civilian form. Two features of these instruments were unique in the testing of high-school reading and interpretation of literature: the use of the wide range of objectives (already presented) and the choosing of a central theme for each test. It was hoped to produce, by means of this second innovation, a natural reading situation which would obviate one of the chief weaknesses of previous tests. The themes chosen for the two forms of the test were "The World of Tomorrow" and "The Mission of American Democracy." Reading selections on these themes, found in the majority of modern textbooks for high schools, held the potentialities for the testing of the wide scope of objectives

set up and gave ample opportunity for testing the thought-processes which are integral parts of reading.

Four criteria, indicated by the basic concepts adopted for the test, determined the choice of selections: (1) that they be appropriate in difficulty and interest to the population to be tested, (2) that they contribute to an understanding of the central theme, (3) that they represent the various mediums of literary expression, and (4) that they exemplify high standards of literary excellence. Each of these tests includes one narrative selection, one essay, and two poetic selections. The exclusion of a drama selection was necessitated by the difficulty of finding a meaningful excerpt which would be short enough for the purposes of the test.

The validity of the choice of selections, in terms of student reaction, was determined by a questionnaire given to students who took both forms of the test. Of eighty-eight students replying, twenty-seven found the tests "very interesting"; twenty-four found them "interesting"; twenty-nine regarded them "fairly interesting"; and eight found them "boring."⁶

Step 2. Constructing items based on the chosen selections in accordance with the objectives.—The method of con-

⁵ For analysis of tests see Verna White, "The Construction of a Test for High-School Students in the Reading and Interpretation of Literature," pp. 140-45. Unpublished Doctor's thesis, University of Chicago, Department of Education, 1945.

⁶ It is significant that the fifty-one who found the tests either "interesting" or "very interesting," had a median mark in English of 81, while pupils who believed it only "fairly interesting" had a median mark of 75.1. With the exception of one respondent, the marks of pupils who considered the tests "boring" were below the median.

structing the items from a formulated set of objectives will not be described in this article, as this has been set forth elsewhere.⁷ There was no attempt to make sure that each item measured only one ability because the generalized nature of the test did not demand this preciseness. There was a definite attempt to test as many of the objectives as possible and to have a comparable distribution of items in the two forms of the test. Evidence concerning the validity of individual items was secured in the standardization program and is reported in the sections of this article dealing with the results of that program.

Step 3. Administering the preliminary forms of the tests and revising them according to the findings.—Each form of the test was given to approximately one hundred twelfth-grade students in a local high school to discover points at which confusion might have been introduced because of ambiguities or lack of clarity either in directions or wording, to ascertain the optimum time for administration of the test, to obtain some idea of the correlation of the test results with the teachers' estimates of the students' abilities, and to obtain a measure of the discriminatory power of each item of the test when that was compared with the performance of the students on the entire test.

The most significant figures from this administration were those which indicated the discriminatory power of

the individual items of the test. On Form B the average index of discrimination per item for 113 items was 22.12, while the median index was 21.00. This figure was derived by dividing the 95 papers into three groups (high, middle, and low) according to total scores and then computing the difference between the percentage of students of the low and the high groups who chose the correct answers. This procedure is recognized as adequately valid for determining discriminatory power.⁸ As a result of the analysis and the subsequent revision, fourteen items were dropped. Some of the less discriminating items were retained but revised. A recalculation of the index of discrimination showed that the average per item was raised to 27 and the median to 24 as a result of dropping the fourteen nondiscriminating items.

Step 4. Administering the revised tryout forms and analyzing the results.—The second tryout Forms A and B were given to 336 and 406 students, respectively, of Grades X–XII, inclusive, over a wide geographical area to discover the power of the test to discriminate among the three grade levels of performance, to gain further evidence on the correlation of test results with the teachers' estimates of the abilities of the students, and to discover the power of each item to dis-

⁷ (a) Verna White and J. B. Enochs, *op. cit.*, pp. 175–76; (b) Verna White, *op. cit.*, pp. 72–74.

⁸ *The Construction and Use of Achievement Examinations*, pp. 39–50. Edited by Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann. Prepared under the Auspices of a Committee of the American Council on Education. Boston: Houghton Mifflin Co., 1936.

criminate between the high- and the low-scoring papers.

Table 1 presents data evidencing the power of the two forms of the test to discriminate between Grades X and XI, Grades XI and XII, and Grades X and XII. It will be noted that there is a statistically significant difference between the means of Grades X and XI and again of Grades XI and XII on each form,⁹ as well as between Grades X and XII. From these figures it may be safely concluded that the tests actually discriminate among students of the three high-school years.

When the schools were asked to participate in the tryout program, the teachers were requested to submit their estimates of the students' abilities in reading and interpreting literature. Table 2 shows the extent of agreement between the teachers' ratings and the marks on the test. It is considered that there was agreement if the test scores of those classed in the superior group fell above the mean and those classed in the poor group fell below the mean. This criterion is recognized to be a very rough estimate, but the percentages would indicate that there was fair agreement and that the test distinguished significantly between the two ends of the scale of achievement.

The discriminating power of each

⁹ This may be concluded from the principle that, if the difference of the two means is four times the probable error of this difference, the difference may be considered statistically significant and that it is caused by a true difference in the two populations being measured.

item was again figured in the same way as it had been for the first form of the test, by computing the difference between the percentage of high and the percentage of low students (according to total test scores) answering each item correctly. For Form B of the test the item analysis after the second tryout showed a discrimi-

TABLE 1

STATISTICAL DATA, FROM EXPERIMENTAL ADMINISTRATION OF ENGLISH TEST TO HIGH-SCHOOL STUDENTS, SHOWING POWER OF TEST TO DISCRIMINATE TENTH-, ELEVENTH-, AND TWELFTH-GRADE PERFORMANCE

Grades Compared	Difference in Means	Probable Error of the Difference	Critical Ratio
Form A:			
X and XI.....	7.20	1.36	5.29
XI and XII.....	3.77	.89	4.23
X and XII.....	10.97	1.15	9.54
Form B:			
X and XI.....	8.52	1.27	6.71
XI and XII.....	6.01	1.18	5.09
X and XII.....	14.53	1.18	12.31

nation of 31 when administered to the three grades. The inclusion of Grades X and XI automatically raised the discriminatory power per item.

Another administration of the second tryout forms was conducted in a high school in Massachusetts and another in California. Both forms of the test were given to the same pupils to obtain a measure of the reliability of the test, derived from the comparability of the two forms. In the school in Massachusetts a reliability coefficient of $.781 \pm .027$ was obtained and

in California, a coefficient of $.877 \pm .02$. That a homogeneously high group had been chosen in Massachusetts was indicated by a mean intelligence quotient of 114.60, with a standard deviation of 10.91. In California the mean intelligence quotient of the group chosen was 104.90, with a standard deviation of 17.14. Therefore the reliability coefficient of

group of students for the testing. The students in the private college did not achieve so high as they might if this testing had not coincided with preparation for college examinations. Despite these unfavorable factors, the critical ratios showed statistically significant differences in all the comparisons between twelfth- and thirteenth-grade scores.

TABLE 2
AGREEMENT OF TEST RESULTS WITH TEACHERS' RATINGS
OF HIGH-SCHOOL PUPILS' ABILITIES

GRADE AND AGREE- MENT ON TEST	FORM A				FORM B			
	Pupils Rated Superior		Pupils Rated Poor		Pupils Rated Superior		Pupils Rated Poor	
	Number	Per Cent	Number	Per Cent	Number	Per Cent	Number	Per Cent
Grade X:								
Agree.....	28	88	4	80	20	91	4	80
Disagree.....	4	12	1	20	2	9	1	20
Grade XI:								
Agree.....	22	73	3	75	29	94	20	87
Disagree.....	8	27	1	25	2	6	3	13
Grade VII:								
Agree.....	55	92	10	50	36	73	20	95
Disagree.....	5	8	10	50	13	17	1	5

$.877 \pm .02$ is probably the more accurate and suggests satisfactory reliability for the test.

In a third type of administration, the second tryout forms were given to Freshmen in three higher institutions to discover whether there was a significant difference in performance between Grades XII and XIII. The results are presented in Table 3. The high school in Massachusetts is the same school mentioned above as having selected a homogeneously good

Step 5. Submitting the revised form to the consultants for criticism and revising the form again where necessary.—Simultaneously with the administration of these tryout tests in the various school populations, the consultants were reviewing and criticizing the tests. The criticisms sent in by the consultants, for the most part, showed general approval of the form and the construction of the test. The most general criticism was that the level of difficulty seemed high. The con-

sultants gave valuable help in suggesting ways of eliminating ambiguities in individual items. They gave suggestions concerning ways in which the administration might be facilitated. They checked the key. The revision of the various items and of the directions took into consideration all the evidence collected from the administrations of the test and from the criticisms of the consultants.

was, rather, to indicate, by the mean performance of a large representative unbiased sample of the country-wide population, a reasonable expectation of achievement on the test. This might be used as the basis for awarding credit to members of the armed forces in the absence of other norms. The comparable civilian form of the test was constructed so that a high school might administer Form B to

TABLE 3
MEAN SCORES ON ENGLISH TEST MADE BY STUDENTS
IN GRADES XII AND XIII

Form and Institution	Number of Students	Mean of Test	Standard Deviation	Probable Error of Mean
Form A:				
A city university.....	79	63.78	10.90	0.007
A state university.....	71	65.26	7.96	.57
High school in Massachusetts.....	89	57.66	12.76	.91
High School I in California.....	73	47.50	14.64	1.16
High School II in California.....	93	50.20	7.92	
Form B:				
A private college.....	115	60.60	11.02	.001
High school in Massachusetts.....	93	57.26	10.37	.73
High School I in California.....	72	48.34	12.88	1.02
High School II in California.....	93	47.05	9.52	

THE STANDARDIZATION PROGRAM

Evidence of validity.—The standardization and the equating of the two forms of the test constituted the final step in the processing of the examination. The program called for the administration of at least a thousand tests of Form A (the military form) and three hundred of Form B (the comparable civilian form) to provide a basis for equating. No attempt was made, in this process of standardization, to set sectional norms for the country. The object of the program

its total school population and from that administration obtain standards at the various levels for its own local situation. Then when the returning veteran takes Form A of the test, he may be awarded credit according to the local standard.

The standardization program yielded the results shown in Table 4. Although the differences in the scores of Grade X and XI and the difference in those of Grades XI and XII for all the cases of Form A appear slight, there was a critical ratio of 8.24 for

the difference in the scores of Grades X and XI, 5.88 for the difference in the scores of Grades XI and XII, and 13.16 for the difference in the scores of Grades X and XII. All these ratios indicate that the differences are statistically significant. There was no further study of the power of the test to discriminate between the twelfth- and thirteenth-grade students.

TABLE 4

MEAN SCORES MADE ON FORM A (MILITARY FORM) OF ENGLISH TEST ADMINISTERED TO 1,000 STUDENTS

Grade	Number of Students	Mean Score	Standard Deviation
Form A, Part I (90 items):			
Grade X.....	902	45	12
Grade XI.....	1,120	48	13
Grade XII.....	876	50	12
Form A, Parts I and II (150 items):			
Grade X.....	762	69	18
Grade XI.....	1,019	74	20
Grade XII.....	814	79	19

Because of the way in which the two forms of the test were to be used, every effort was made to insure comparableness. The items were con-

the same literary types. The items were selected by the same method of choosing discriminating items. The two forms of the test were tried out on the same or similar populations. Despite this care in construction, it was necessary to ascertain exactly how comparable these two tests were. For this purpose three hundred of each of the forms were alternated within the same classrooms at each grade level. It is assumed that, by such a process, comparable results will be obtained because the same numbers of pupils from comparable situations are selected and because a random sampling is obtained by selecting every other student within similar groups seated according to no specified plan. This administration yielded a mean score of 71 on Form A (for 796 cases) and 69 on Form B (737 cases). The standard deviation was 18 on both forms. The closeness of the means and the identicalness of the standard deviations indicate comparableness of the two forms. Because of the slight disparity in the means, however, it was necessary to compute equivalent scores by means of the equating formula:

$$B\text{-score expressed in comparable } A\text{-score terms} = \left[\left(\frac{\text{Raw Score}_B - M_B}{S.D._B} \right) S.D._A \right] + M_A$$

structed upon the same objectives, with approximately the same distribution of items per objective. The two forms of the test contained selections of approximately equal difficulty for

With this set of equivalent scores, it is possible to assign any score on Form A a comparable score on Form B. A provision of this kind was necessary to implement the plans for using

the civilian form for establishing local norms to be employed in interpreting veterans' records on Form A.

The fact that the test discriminated among tenth-, eleventh-, and twelfth-grade students was one indication of the validity of the test. The test also showed a consistent power to discriminate between good and poor students at each grade level. The schools which

each of these groups, with the results shown in Table 5. This simple statistical device demonstrates that in each case there is a substantial difference between the medians of those groups judged superior and those judged poor by the teachers. This finding adds definite evidence that the test consistently differentiates between those two categories of stu-

TABLE 5
POWER OF THE TEST TO DISCRIMINATE BETWEEN SUPERIOR AND POOR STUDENTS AT TENTH-, ELEVENTH-, AND TWELFTH-GRADE LEVELS

GRADE AND GROUP	PART I (90 ITEMS)		PARTS I AND II (150 ITEMS)	
	Number of Students	Median Score	Number of Students	Median Score
Grade X:				
Superior students.....	58	51.1	45	82.7
Poor students.....	42	38.7	23	58.0
Grade XI:				
Superior students.....	32	58.36	25	93.0
Poor students.....	22	40.5	31	64.9
Grade XII:				
Superior students.....	23	58.0	23	86.8
Poor students.....	22	33.4	22	48.5

co-operated in the standardization program were asked to give the names of those students whom they considered superior in English and those whom they considered either borderline or failing students. With full recognition of the weakness of correlating test scores with teachers' estimates, a tabulation of the scores obtained by those students considered superior and those considered poor was made for Part I and for the total scores. The medians were found for

dents and thus contributes evidence of the validity of the test.

The most striking proof that items of the test were validly constructed was found in the results of the testing in a Detroit high school. An analysis of individual items showed that the items for which these pupils made the most significant gain from Grade X to Grade XI and from Grade XI to Grade XII were those constructed to measure the following behavior:

- a) He develops a greater sensitivity to man, society, and its institutions.
- b) He responds to the system of values expressed by the author.
- c) He grasps the social significance of the actions and speeches of the characters and the expositions and implications of the author.

A letter of inquiry addressed to Stanley Dimond, director of the Citizenship Education Study, brought the following explanatory facts concerning the Detroit high school in which the tests were given. It was one of the

reading and composition had been considered particularly good agencies for promoting these attitudes and appreciations.

To extract still further evidence from these data, the results from Grade XII in Detroit were compared with those of another midwestern city (designated as "City Y" in this discussion) known to follow a traditional program. When the twelve items in which Detroit showed the greatest superiority over City Y were

TABLE 6
RELIABILITY COEFFICIENTS FOR FORM A (TOTAL TEST) FOR ALL ADMINISTRATIONS IN THE STANDARDIZATION PROGRAM

Grade	Number	Mean Score	Standard Deviation	Reliability Coefficient
All tenth-grade cases.....	762	69	18	0.89
All eleventh-grade cases.....	1,019	74	20	.91
All twelfth-grade cases.....	814	79	19	.90
All grades.....	2,595	74	20	0.91

participating schools in the Michigan Study of the Secondary School Curriculum, had experimented with a core course for some of its students, was one of the largest users of the *Scholastic* magazine of any of the high schools in the United States, and had an active intercultural committee. Further study of the Detroit curriculum¹⁰ showed that there had been a direct frontal attack on the teaching of democratic principles and that

¹⁰ a) Stanley E. Dimond, "Detroit Uses School Committees," *Intercultural Education News*, VI (January, 1945), 1-2.

b) "More of the Detroit Plan," *American Unity*, III (January, 1945), 7-9.

isolated, it was found that seven of the twelve tested the reading abilities promoting greater social understanding. When the twelve items in which City Y showed the greatest superiority over Detroit were isolated, it was found that eight of them measured the ability to identify rhetorical devices or grammatical constructions. Only one of the items in which Detroit was superior tested knowledge of rhetorical devices, and none of the items in which City Y was superior tested reading abilities promoting social understandings. A safe conclusion is that in these two schools of

known emphasizes the test was measuring those things which it purported to measure.

Evidence of reliability.—It was assumed that reliability was introduced into the test at the time of its construction because of the use of a carefully formulated set of objectives as a consistent basis for choosing items, because of the length of the test, and because of the objectivity of the scoring. The figures from the standardization program substantiate this hypothesis. The coefficient of reliability was computed by the split-halves method and then corrected for at-

tenuation by the Spearman-Brown formula.

The results for the total test (150 items) for all the administrations of Form A in the standardization program are given in Table 6. The coefficients justify the assertion that the test measures whatever it is measuring with a degree of reliability which is accepted by authorities as sufficiently high.

Substantially the same results were found when the equating cases were computed for reliability. A coefficient of .90 was found for Form A and .89 for Form B.

MEASURING ACHIEVEMENT IN HIGH-SCHOOL ENGLISH

VERNA WHITE

Reprinted from

The School Review

Published monthly (except July and August) by

The University of Chicago at the University of Chicago Press

Under the Direction of the Editorial Committee of the Department of Education

Subscription price, \$3.00 a year

PRINTED IN THE U.S.A.

MEASURING ACHIEVEMENT IN HIGH-SCHOOL ENGLISH

VERNA WHITE
Syracuse University



AN EARLIER article¹ presented the process by which the English test of the United States Armed Forces Institute was constructed to insure the greatest possible validity and reliability. It also included data concerning the validity and the reliability which were derived from a nation-wide standardization program. The present article discusses the extension of the validity study by means of the case-study method applied to one institution.

Because it was realized that the assumptions involved in the use of large-group statistical procedures for establishing validity and reliability of a test are not necessarily appropriate to a particular school situation and that many factors peculiar to a certain school system and individual students within that school system are ignored in such studies as that described in the previous article, a study was made of the performance of the twelfth-grade students in a single city school system in the Middle West. The investigation of the results of the English test was conducted in two parts to give further evidence of validity.

¹ Verna White, "Measuring Competence in English of Armed Services Personnel," *School Review*, LV (June, 1947), 345-55.

In the first step of the first part, the performance of the whole twelfth-grade group (344 students) was studied in light of the facts gathered about the school and its program. The belief was that the test could be considered valid if it distinguished among students of different curriculums according to the varying richness of those curriculums; if the correlation between the scores on the test and intelligence-test scores was approximately .65; and if the majority of students received correct scores on those items which were in harmony with the objectives of the school and of the English course and, conversely, if smaller proportions of students made correct scores on those items not related to the objectives of the school.

In the second step of the first part, the performance of the students (sixty-four in number) who selected English in their twelfth year was studied. The test could be considered valid if the correlation between the English test and intelligence-test scores for the selected group of students was substantially lower than that for the twelfth-grade group as a whole; if it showed a difference in performance between the students of

the college-preparatory course taking English in their twelfth year and those in the same course not taking English in their twelfth year; and if it showed a similarity in performance of this group with that of a similar group in another school that followed a similar program.

In the second part the performance of the students who selected English in their twelfth year was given intensive study, through use of the data concerning the whole school program and, in addition, personal data from their records from Grade VI through XII. It was believed that there would be further evidence of validity if the test distinguished between students who had had consistently good records in English and those who had had consistently poor records and if logical explanations could be found for marked discrepancies between performance on the test and the performance expected from an estimate by the teacher.

The first procedure was to secure an acquaintance with the various phases of the school program through visiting classes; discussing the work informally with the teachers, the principal, and guidance officers; examining textbooks, notebooks, term papers, book reviews, and book lists; visiting the library and discussing with the librarian the policies governing library administration and practice; examining the prescribed curriculums for the various high-school years; and chatting informally with the students.

The general philosophy of the school

was traditional. The emphasis was placed on assigned work, with little apparent co-operative teacher-pupil planning. The curriculum was based on a series of courses which were entities within themselves, and there was little attempt at integration or correlation. The teaching, for the most part, was subject-centered. The personnel program appeared detached from the rest of the school program, and the testing program did little to facilitate diagnosis, placement, or curriculum improvement in the English program. The school drew only slightly on the available rich community resources. The courses of study of the various academic divisions contained the usual requirements, and scant attention was given to individual differences. There was, however, a library adequate to meet the needs of an extensive reading program, and the teacher of Grade XII was conducting a variation of the prescribed program to give the students experience in reading and analyzing contemporary books. The students demonstrated their ability to read and organize the pertinent material. In these classes, freedom of activity and pupil direction, which were not observed in the majority of the classes in the school, were apparent.

STUDY OF ALL SENIORS

In Table 1 are reported the results of the testing of the entire twelfth-grade population according to the courses of study pursued. The median intelligence quotient is given for each

group because it appears to be a factor in the difference in scores and also reflects the traditional policy of assigning the students of low intelligence to the general course. The table presents results for both sections of the test and

the use of these data in further statistical study, the following conclusions concerning the validity of the test were drawn.

The test proved to have the power to discriminate among students in

TABLE 1
RESULTS ON TWO PARTS OF ENGLISH TEST TAKEN BY 344 TWELFTH-GRADE
STUDENTS ENROLLED IN FOUR HIGH-SCHOOL COURSES

Part of Test and High-School Course	Number of Students	Median Intelligence Quotient	Median Score*	Mean	Standard Deviation	Probable Error of Mean
Part I. Reading and Interpretation of Literature:						
College-preparatory.....	162	105.1	51.38	50.93	11.06	0.586
General.....	85	93.2	36.18	34.92	11.90	.870
Commercial.....	71	99.5	41.12	41.86	11.20	.897
Industrial.....	26	94.7	36.50	37.46	9.87	1.305
Total.....	344	100.7	43.61	44.08	13.12	
Part II. Literary Acquaintance:						
College-preparatory.....			31.33	30.70	6.50	0.344
General.....			22.96	22.95	5.92	.433
Commercial.....			23.12	23.01	6.33	.507
Industrial.....			20.50	21.12	6.45	.853
Total.....			25.73	26.50	7.58	
Parts I and II:						
College-preparatory.....			80.95	81.67	16.17	0.857
General.....			58.55	57.98	16.04	1.174
Commercial.....			63.35	65.08	15.08	1.207
Industrial.....			57.50	58.84	14.28	1.889
Total.....			69.20	70.64	18.93	

* Possible scores: Part I, 90; Part II, 60; Parts I and II, 150.

for the composite scores, although the major emphasis of the study was on Part I of the test.

When the scores of the whole Senior class, the college-preparatory course, and the general course were correlated with the intelligence quotients, the coefficients presented in Table 2 were obtained.

With the figures given above and

courses of study varying in richness and among students of varying intelligence quotients. The differences in intelligence quotients and in scores for the college-preparatory and general groups are clearly apparent. The critical ratios for the differences in means of these two groups were 15.25 for Part I and 16.34 for Parts I and II —ratios which would indicate that

there was no question concerning the statistical significance of the differences.

To determine whether all this difference could be accounted for by difference in intelligence, the scores of 50 pairs of students (paired according to intelligence quotient) from the college and the general courses were correlated, the critical ratio computed, and the means of the two groups compared. In the reading and interpretation of literature, there was a critical ratio of 4.52, with a mean score of 45.80 for the college group and 38.66 for the general group. For the total scores in reading and interpretation of literature and in literary acquaintance, there was a critical ratio of 7.35 with a difference of means of 15.0 (a mean of 77.5 for the college group and 62.5 for the general). It seems clear that the program of the general course was not producing results in accordance with the intellectual capacities of the students.

When the scores of 59 pairs of Seniors (paired according to intelligence quotient) taking English and those of Seniors not taking English were compared, no statistically significant difference was found in the scores for the reading and interpretation of literature, but there was a critical ratio of 4.52 in the literary-acquaintance section of the test, with a mean of 31.91 for the English group and 27.64 for the non-English. This would seem to give some evidence, though slight, concerning the validity of the test, for these students were given the same

program by the same teachers through Grades X and XI, when the power to read and interpret literature was being developed. The groups of Seniors who were taking English were doing so optionally and would, therefore, be the students most interested in literary study and wide reading. One of the main objectives of the twelfth-grade English program was to extend acquaintance with books and authors,

TABLE 2
CORRELATION OF INTELLIGENCE QUOTIENTS
WITH SCORES ON TWO PARTS OF ENGLISH
TEST FOR ALL SENIORS AND STUDENTS IN
COLLEGE-PREPARATORY AND GENERAL
COURSES

GROUPS OF STUDENTS	CORRELATION	
	Part I	Parts I and II
All Seniors.....	.682	.694
College-preparatory....	.550	.578
General.....	.528	.486

both in class and through an extensive reading program.

As another phase of the study of the validity of the test, 72 students from the total Senior class were paired according to intelligence quotient with 72 students from another city system of approximately the same size. Since both these school systems follow the traditional high-school English program, one would expect the results to be similar. The means of the 72 cases which were paired were 56.6 for the other school system and 53.0 for the school being studied. The critical ratio of 1.78 indicates no statistically

significant difference between the performance of the two groups.

The correlation of .68 between intelligence quotients and scores on Part I of the test for the 344 Seniors is indicative of the validity of the test, since a good reading test should correlate at about .65 with a measurement of intelligence. The lower correlation of the intelligence quotients and test scores for the college-preparatory group, a more homogeneous group, is another indication that the factors being measured operate according to statistical principle; that is, the more homogeneous the group, the lower the correlation. The even lower correlation of the general group is significant in light of the evidence that the general group was not achieving as effectively from its course of study in relation to its intelligence quotient as was the college-preparatory group.

Another indication of the validity of the test was presented in the previous article. The results from this city school system were compared with the results of a Detroit high school, a system noted for its teaching of social understandings, to discover whether differences in percentages of correct responses for the individual items followed any particular pattern in harmony with the facts known about the two curriculums. It was discovered that the students of this school, known to follow a traditional program, showed their greatest superiority in the ability to identify rhetorical devices and grammatical constructions and their greatest weakness in

those items testing social understandings, while the students of the Detroit school demonstrated superiority on items designed to test social understandings and their greatest weakness in the identification of rhetorical devices and grammatical constructions. It was a clear-cut demonstration of competence in those areas stressed in the teaching of the two schools.

From the analysis of the results of the administration of the test in the school under study the following evidences of validity were found: (1) The test discriminated among students of varying intelligence, as was shown through a correlation of .68 between the test scores and intelligence quotients. (2) The correlations between intelligence quotients and test scores followed the expected pattern. (3) The test discriminated among students of the same intelligence quotients in varying curriculums. (4) The test showed no statistically significant difference between the performance of the college-preparatory group taking English and those pupils not taking English, except in the area of literary acquaintance. (5) The test revealed no statistically significant difference between the scores of students of two high schools both of which teach the traditional type of college-preparatory course. (6) Through the item analysis the test identified those strengths and weaknesses which one would expect from the school program.

When the odd-even method of finding reliability was used and the Spearman-Brown formula for attenuation

was applied, the resulting coefficient was .888. The means of the two halves were 22.52 and 21.67, with standard deviations of 6.98 and 7.06, respectively. According to statistical findings, a coefficient above .88 for a test given within a single grade is considered sufficiently reliable.²

INTENSIVE STUDY OF STUDENTS WHO ELECTED ENGLISH

To validate still further the power of the test to distinguish among various types of students within a more homogeneous group, an intensive study was made of the performance of the sixty-four twelfth-grade students who were taking English. For these students, facts were collected concerning their intelligence quotients and the teacher's estimate of their ability in English. In addition, data were assembled in the following areas: age; general rating by all faculty members in intellectual capacity, intellectual achievement, initiative, integrity, emotional control, social attitudes, etc; number of siblings in family; financial situation; home situation; special abilities; interests; hobbies; clubs; awards; academic record from Grade VI through Grade XII; personality data; health data; work done outside of school; college and vocational choices; books read in Grades XI and XII; comments on final term papers and notebooks; etc.

² Harry A. Greene and Albert N. Jorgensen, *The Use and Interpretation of High School Tests*, p. 111. New York: Longmans, Green & Co., 1936.

It was believed that abnormalities in any of these areas would affect the ability to achieve on the test more vitally than would the intelligence quotient, as all the students were within the normal range, and that the failure of some students to achieve at expected levels was not evidence of lack of validity of the test. With this major premise in mind, it was hypothesized that the test could be considered valid if (1) a prediction of performance on the test could be made from an examination of the personal data by judges who understood the objectives of the examination; (2) discrepancies between expectation of performance based on intelligence quotient and the teacher's estimates and achievement on the test could be explained logically by the other data collected; (3) students who had been judged as having good critical powers achieved well, and vice versa; (4) students having optimum adjustment achieved better than those who did not; (5) students who had had a good record in literary activities and who demonstrated a current interest in literary activities achieved well on the test, and vice versa.

Two co-workers in the testing of English who were familiar with the test and the objectives upon which it was constructed were asked to examine the personal data and classify the sixty-four students into the top quarter, the middle two quarters, and the low quarter, according to expected achievement on the total test. They placed in the top quarter thirteen of

the seventeen who achieved in the top quarter on the test; in the middle two quarters, twenty-one of the thirty who achieved in the middle two quarters on the test; and in the low quarter, twelve of the seventeen who achieved in the low quarter on the test. They predicted within the correct quarter 76 per cent of the top quarter, 70 per cent of the middle two quarters, and 71 per cent of the low quarter. No student placed in the high quarter by the judges fell in the low quarter of achievement. Likewise, no student placed low fell in the top quarter. This result gave some verification to the belief that the test was validly testing those objectives for which it was constructed.

When the test was given, the teacher was asked to select the students with the greatest ability in reading and interpretation of literature and those with the lowest ability. Of eleven students assigned by her to the top quarter six (55 per cent) achieved in the top quarter in the total test; of 10 students assigned to the lowest quarter six (60 per cent) achieved in the low quarter on the total test. All students adjudged of highest ability by the teacher fell above the median in test performance; of those students adjudged of lowest ability by the teacher three fell above the median. A few representative examples of those students whose performance did not conform with the teacher's estimate are cited below. In each case, factors are presented from the supplementary data which suggest either an under-

estimate or an overestimate by the teacher.

Pupil A.—A girl with an intelligence quotient of 110 with general ratings of "excellent" and "good" by the teachers was rated in the top quarter by the teacher but did not achieve in the top quarter. She listed no special interests and named piano-playing and dancing as her special abilities.

The only club to which she belonged was a temperance group. The anecdotal comments from the records show that she was an excellent reader in Grade VI, but there is no mention beyond that of excellence in the subject. It was evident from the final paper that she was passed because she did careful work that a teacher would appreciate, but her paucity of interests militated against the best performance on the test of reading and interpretation of literature.

Pupil B.—A boy with an intelligence quotient of 118 and general ratings of "excellent" by the teachers was not included in the top quarter by the English teacher but achieved first place in the total score. This boy listed as special abilities: languages, history, public speaking, and music. His interests were stamps, coins, piano-playing, speaking, cards, Eagle Scout, and sports. Clubs to which he belonged were Spanish, Pan-American, political science, school-paper staff, astronomy, and church choir. He had received the Rotary debating award and had chosen law as his profession. In Grade X he was rated as "an excellent student in English," and in Grade XI the teacher noted, "He is something of a genius in English, best student I ever had, a clear thinker and analyzer." That he was not rated at the top by the teacher might well be due to the fact that, owing to the diversification of his interests, he did not do such spectacular daily work as some pupils whose interests were primarily literary. He did not read a large number of books, and those that he did read were of a mature type on subjects in line with his interests.

Pupil C.—A boy with an intelligence quotient of 107 and general ratings of "average" and "good" achieved a score in the bottom quarter but was not rated there by the English teacher. As his special abilities he listed swimming and baseball, and music and stamps were his special interests. He belonged to no clubs. The comments of the teachers concerning his abilities were conflicting. In Grade X various teachers say about him; "poor student," "likes to get something for nothing by way of marks," "pleasant," "low ability," "did excellent work in English." In Grade XI his English teacher said, "Has done very poor work for me in 11-2 English, because he will not work. He seems very immature and childish in mind." His mathematics teacher believes "there is something wrong with his I.Q. rating or else something unknown to us holds him back. I believe he tries but does not accomplish in mathematics." An essay written by this student during the last week of school on "The Importance of Having Someone to Whom to Explode" is very revealing in the light of these comments. It is an account of the trials of an adolescent boy who is never given any credit in the home for the things that he considers he has done well and with every intention of pleasing. Here, it would seem, was a case of an adolescent attempting to meet tasks of development and finding considerable emotional blocking from his failures. His special abilities and interests did not indicate motivation in literary pursuits.

These examples are illustrative of those cases in which there was a discrepancy between the achievement on the test and the teacher's rating of the student in the four quarters. The weakness of using teachers' estimates as a criterion against which to validate a test is well demonstrated by these findings.

In addition to the study of individ-

ual cases to justify the discrepancy between expected and actual performance, the students were studied as groups, according to scores on the test, for the purpose of identifying characteristics of each group and of discovering common characteristics which made for success or lack of success. The divisions of top quarter, middle two quarters, and bottom quarter were also employed for this part of the study. It is well to keep in mind, when the characteristics of these three groups are considered, the homogeneity of the intelligence quotients. The median was 106, with an interquartile range of 102-115 and a spread from 93 to 121. There were certain factors outside the intelligence factor which recurred in one group and either did not occur at all, or occurred to a lesser degree, in the other groups.

The findings of significance are summarized as follows: (1) The group of students in the top quarter had an all-round record of excellence in achievement; the middle group achieved well; and the bottom quarter did average work. (2) The top quarter was a well-adjusted group of students; the middle group had some emotional difficulties of a more or less serious nature; and a large proportion of the students in the bottom quarter indicated serious maladjustments. (3) The number of interests, clubs, and hobbies indicated by the students in the anecdotal records was about the same for each of the groups. The top quarter showed a high number of literary and dramatic interests and a low number of social and

sports interests; the middle and bottom groups showed high numbers of interests in sports and music. (4) In the vocational choices the students in the top quarter indicated professions as their choices; the middle group showed a greater number of choices, with some professions mentioned; and the bottom quarter, for the most part, avoided the professional type of work. (5) Twenty-nine per cent of the students in the top quarter were working; 33.3 per cent of the middle group; and 47.0 per cent of the bottom quarter.

Although all the factors mentioned above are significant, it would appear that the emotional adjustment of the student was the one single factor in this study which skewed results most. Because of this, it seems worth while to set forth some of these emotional problems in the words of the anecdotal records of the teachers. As was stated above, there were no emotional difficulties mentioned for students in the top quarter. In the case of students in the middle group, there were some serious difficulties and some not so serious. Some of the difficulties mentioned for the middle group are:

Too dependent on mother.
Has locked herself within self, shy, retiring.
Lazy—no initiative.
Retiring, daydreaming, lazy.
Seems to lack emotional stability.
Doesn't care about exerting himself.
Poor home conditions, much absence, tired from overwork.
Strictly religious environment, not enough social life.
No school clubs or social activities, nervous condition, apt to have a chip on her shoulder.

Too quiet, hard to draw out.

No school clubs, extremely sensitive, frightened.

World against her, wants to reform all, very quiet, poor attitude.

For pupils in the bottom group the following difficulties were mentioned:

Great feelings of doubt, shy, demure, always alone.

Likes to get something for nothing by way of grades, seems very immature, childish, something wrong which holds him back.

Little immature, mother has determination to make a student of him.

Forlorn, undernourished, retiring, shy, too quiet for his own good, needs to grow up and accept responsibility, no clubs, good citizen in study hall.

Weak sister, physical capacity low, rather emotionally unstable, absent a good deal, lives with guardian, father a doctor in near-by town.

Broken home, stepsister born in '42; needs to come out of her shell, always late, shy, no social life.

Tries to be funny constantly, boisterous, obese, too easygoing, seems immature; no clubs or social activities.

Mother died in February of this year, student has a heart condition.

Needs a jolt by someone, moody and despondent, handicapped by a health condition which exerts pressure on her head, causing forgetfulness and lack of alertness at times; hard for her to adjust to requirements and standards.

The intensive case-study technique yielded further evidence of the validity of the test as a measure of the ability to read and interpret literature. It demonstrated that discrepancies occurring between expected performance based on teachers' estimates and the performance on the test are not necessarily indications of a lack of

validity of the test but may be accounted for by other factors influencing the score. It showed that these personal data form a reliable body of facts for predicting performance on the test when the objectives upon which the test is constructed are thoroughly known. It demonstrated that

the test has the power to classify students into high, low, and medium groups according to generalized characteristics. It has shown the necessity of supplementing large-scale standardizing programs with a study of personal data if the true validity of the test is to be ascertained.

